Automatisierung in der Kataloganreicherung

A. Wagner

alexander.wagner@tu-ilmenau.de

UNIVERSITÄTSBIBLIOTHEK ILMENAU

Überblick

- Motivation
- Anforderungen
- Erfassung
- Texterkennung
- Upload
- Prüfung
- Zusammenfassung

Motivation

UNIVERSITÄTSBIBLIOTHEK ILMENAU

Für Benutzer:

Motivation

UNIVERSITÄTSBIBLIOTHEK ILMENAU

Für Benutzer:

- Suchbar
- Indexiert
- Recherchierbar

Motivation

UNIVERSITÄTSBIBLIOTHEK ILMENAU

Für Benutzer:

- Suchbar
- Indexiert
- Recherchierbar

Für Bibliothek:

Motivation

UNIVERSITÄTSBIBLIOTHEK ILMENAU

Für Benutzer:

- Suchbar
- Indexiert
- Recherchierbar

Für Bibliothek:

- Vereinfachung
- Beschleunigung
- Automatisierung:

Motivation

UNIVERSITÄTSBIBLIOTHEK ILMENAU

Für Benutzer:

- Suchbar
- Indexiert
- Recherchierbar

Für Bibliothek:

- Vereinfachung
- Beschleunigung
- Automatisierung:
  - FTP-Transfer
  - Linkcheck
  - Transferkontrolle

Früher

1. TOC-Scan als *PDF* (bei Katalogisierung)
2. Eintrag TOC-Link im Katalog (Makro)
3. *Manueller* Upload
4. *Manueller* Linkcheck
5. *Ein* Arbeitsplatz mit Adobe Acrobat (Korrekturen)

Anforderungen

- Hohe Erkennungsrate **(keine manuelle Kontrolle)**

Anforderungen

- Hohe Erkennungsrate (keine manuelle Kontrolle)
- Mehrsprachige Texte (deutsch, englisch)

Anforderungen

UNIVERSITÄTSBIBLIOTHEK ILMENAU

- Hohe Erkennungsrate (keine manuelle Kontrolle)
- Mehrsprachige Texte (deutsch, englisch)
- Nachträgliche Texterkennung (OCR)

Anforderungen

UNIVERSITÄTSBIBLIOTHEK ILMENAU

- Hohe Erkennungsrate (keine manuelle Kontrolle)
- Mehrsprachige Texte (deutsch, englisch)
- Nachträgliche Texterkennung (OCR)
- Flexible Integration (Geschäftsgang)

Anforderungen

UNIVERSITÄTSBIBLIOTHEK ILMENAU

- Hohe Erkennungsrate (keine manuelle Kontrolle)
- Mehrsprachige Texte (deutsch, englisch)
- Nachträgliche Texterkennung (OCR)
- Flexible Integration (Geschäftsgang)
- *Akzeptanz* (Katalogisierung)

Anforderungen

UNIVERSITÄTSBIBLIOTHEK ILMENAU

- Hohe Erkennungsrate (keine manuelle Kontrolle)
- Mehrsprachige Texte (deutsch, englisch)
- Nachträgliche Texterkennung (OCR)
- Flexible Integration (Geschäftsgang)
- *Akzeptanz* (Katalogisierung)
  - Einfache Erfassung
  - Geringer Zeitbedarf
  - Fehlerprüfung

Anforderungen

UNIVERSITÄTSBIBLIOTHEK ILMENAU

- Hohe Erkennungsrate (keine manuelle Kontrolle)
- Mehrsprachige Texte (deutsch, englisch)
- Nachträgliche Texterkennung (OCR)
- Flexible Integration (Geschäftsgang)
- *Akzeptanz* (Katalogisierung)
  - Einfache Erfassung
  - Geringer Zeitbedarf
  - Fehlerprüfung
- *Automatisierung*

Anforderungen

- Hohe Erkennungsrate (keine manuelle Kontrolle)
- Mehrsprachige Texte (deutsch, englisch)
- Nachträgliche Texterkennung (OCR)
- Flexible Integration (Geschäftsgang)
- *Akzeptanz* (Katalogisierung)
  - Einfache Erfassung
  - Geringer Zeitbedarf
  - Fehlerprüfung
- *Automatisierung*
  - Texterkennung
  - Upload
  - Linkcheck
  - Transfervergleich

Allgemeines

UNIVERSITÄTSBIBLIOTHEK ILMENAU

- Sprache: *Perl*

Allgemeines

- Sprache: *Perl*
- *Portabel* (plattformunabhänig)

Allgemeines

- Sprache: *Perl*
- *Portabel* (plattformunabhänig)
- Kleine Einzelscripte (flexibel)

Allgemeines

- Sprache: *Perl*
- *Portabel* (plattformunabhänig)
- Kleine Einzelscripte (flexibel)
- Konfiguration über Variablen

Allgemeines

- Sprache: *Perl*
- *Portabel* (plattformunabhänig)
- Kleine Einzelscripte (flexibel)
- Konfiguration über Variablen
- Arbeit zwischen zwei Verzeichnissen

Erfassung

UNIVERSITÄTSBIBLIOTHEK ILMENAU

- Scanner:

Erfassung

- Scanner:
  - Leise (aber: Schrittmotor)
  - Zügig
  - Preiswert
  - Einfache Software (ggf. `sane`)

Erfassung

UNIVERSITÄTSBIBLIOTHEK ILMENAU

- Scanner:
  - Leise (aber: Schrittmotor)
  - Zügig
  - Preiswert
  - Einfache Software (ggf. `sane`)

A4-USB-Flachbettscanner

Erfassung

UNIVERSITÄTSBIBLIOTHEK ILMENAU

- Scanner:
  - Leise (aber: Schrittmotor)
  - Zügig
  - Preiswert
  - Einfache Software (ggf. `sane`)

  A4-USB-Flachbettscanner
- Single-Page-Scan (Tiff, 400dpi, Graustufen)

Erfassung

- Scanner:
  - Leise (aber: Schrittmotor)
  - Zügig
  - Preiswert
  - Einfache Software (ggf. `sane`)

  A4-USB-Flachbettscanner
- Single-Page-Scan (Tiff, 400dpi, Graustufen)
- Fortlaufende Numerierung (Dateiname = ppn + Nummer)

Erfassung

UNIVERSITÄTSBIBLIOTHEK ILMENAU

- Scanner:
  - Leise (aber: Schrittmotor)
  - Zügig
  - Preiswert
  - Einfache Software (ggf. `sane`)

  A4-USB-Flachbettscanner
- Single-Page-Scan (Tiff, 400dpi, Graustufen)
- Fortlaufende Numerierung (Dateiname = ppn + Nummer)
- Einzelseiten: gemeinsamem Verzeichnis (LAN-Share)

Erfassung: Vorteile

- Zeitbedarf für Scan (gegenüber PDF)

Erfassung: Vorteile

- Zeitbedarf für Scan (gegenüber PDF)
- Leiser (Schrittmotor: empfunden)

Erfassung: Vorteile

- Zeitbedarf für Scan (gegenüber PDF)
- Leiser (Schrittmotor: empfunden)
- Bessere Scanqualität

Erfassung: Vorteile

UNIVERSITÄTSBIBLIOTHEK ILMENAU

- Zeitbedarf für Scan **(gegenüber PDF)**
- Leiser **(Schrittmotor: empfunden)**
- Bessere Scanqualität
- Einfache Korrektur **(kein Adobe Acrobat nötig)**

Erfassung: Vorteile

UNIVERSITÄTSBIBLIOTHEK ILMENAU

- Zeitbedarf für Scan **(gegenüber PDF)**
- Leiser **(Schrittmotor: empfunden)**
- Bessere Scanqualität
- Einfache Korrektur **(kein Adobe Acrobat nötig)**
- OCR unabhänig vom Scanvorgang

Erfassung: Vorteile

UNIVERSITÄTSBIBLIOTHEK ILMENAU

- Zeitbedarf für Scan **(gegenüber PDF)**
- Leiser **(Schrittmotor: empfunden)**
- Bessere Scanqualität
- Einfache Korrektur **(kein Adobe Acrobat nötig)**
- OCR unabhänig vom Scanvorgang
- Signifikant höhere Erkennungsraten

Erfassung: Vorteile

UNIVERSITÄTSBIBLIOTHEK ILMENAU

- Zeitbedarf für Scan (gegenüber PDF)
- Leiser (Schrittmotor: empfunden)
- Bessere Scanqualität
- Einfache Korrektur (kein Adobe Acrobat nötig)
- OCR unabhänig vom Scanvorgang
- Signifikant höhere Erkennungsraten
- Nachbearbeitung bereits erfasster Dateien (transparent)

Erfassung: Vorteile

UNIVERSITÄTSBIBLIOTHEK ILMENAU

- Zeitbedarf für Scan (gegenüber PDF)
- Leiser (Schrittmotor: empfunden)
- Bessere Scanqualität
- Einfache Korrektur (kein Adobe Acrobat nötig)
- OCR unabhänig vom Scanvorgang
- Signifikant höhere Erkennungsraten
- Nachbearbeitung bereits erfasster Dateien (transparent)
- Flexibel im Geschäftsgang anzuordnen

Erfassung: Nachteile

- Viele Dateien **(Jura: z. T. > 40 Dateien/Buch)**
- Temporär hoher Platzbedarf **(Tiff)**

OCR

Vorarbeit: „Buchbinder” – **createbook.pl**

OCR

UNIVERSITÄTSBIBLIOTHEK ILMENAU

Vorarbeit: „Buchbinder” – **createbook.pl**

Abby FineReader 8 Professional

OCR

Vorarbeit: „Buchbinder” – **createbook.pl**

Abby FineReader 8 Professional

- Preiswert
- Hohe Erkennungsrate
- Erzeugung von Multilayer-PDF **(Bild über Text)**
- Unabhängig vom Eingansgformat **(Nachbearbeitung alter PDFs)**
- Automatische Seitenrotation **(Buchscanner)**

OCR: Schwierigkeiten

UNIVERSITÄTSBIBLIOTHEK ILMENAU

- Sonderzeichen (z. B. Mathematik)
- Speicherverwaltung FineReader (Windows-spezifisch)
- Halbautomatisch (Windows-spezifisch)

OCR: Arbeitsschritte

UNIVERSITÄTSBIBLIOTHEK ILMENAU

1. Eingangsverzeichnis (LAN):

   <ppn>.tif, <ppn>0001.tif, <ppn>0002.tif

OCR: Arbeitsschritte

UNIVERSITÄTSBIBLIOTHEK ILMENAU

1. Eingangsverzeichnis (LAN):
   <ppn>.tif, <ppn>0001.tif, <ppn>0002.tif
2. Multipage-Tiff (OCR-Input, local)

OCR: Arbeitsschritte

UNIVERSITÄTSBIBLIOTHEK ILMENAU

1. Eingangsverzeichnis (LAN):
   <ppn>.tif, <ppn>0001.tif, <ppn>0002.tif
2. Multipage-Tiff (OCR-Input, local)
   - *Ziel: ein Tiff pro ppn*
   - Extraktion aller ppn (Dateiname)
   - Zusammenfügen: **tiffcp**
   - **createbook.pl**

OCR: Arbeitsschritte

1. Eingangsverzeichnis (LAN):
   <ppn>.tif, <ppn>0001.tif, <ppn>0002.tif
2. Multipage-Tiff (OCR-Input, local)
   - *Ziel: ein Tiff pro ppn*
   - Extraktion aller ppn (Dateiname)
   - Zusammenfügen: **tiffcp**
   - **createbook.pl**
3. **FineReader**-Job (manuell, local):

OCR: Arbeitsschritte

UNIVERSITÄTSBIBLIOTHEK ILMENAU

1. Eingangsverzeichnis (LAN):
   <ppn>.tif, <ppn>0001.tif, <ppn>0002.tif
2. Multipage-Tiff (OCR-Input, local)
   - *Ziel: ein Tiff pro ppn*
   - Extraktion aller ppn (Dateiname)
   - Zusammenfügen: **tiffcp**
   - **createbook.pl**
3. **FineReader**-Job (manuell, local):
   - Lese OCR-Input
   - Erzeuge Doublelayer-PDF (Upload)

OCR: Arbeitsschritte

UNIVERSITÄTSBIBLIOTHEK ILMENAU

1. Eingangsverzeichnis (LAN):
   <ppn>.tif, <ppn>0001.tif, <ppn>0002.tif
2. Multipage-Tiff (OCR-Input, local)
   - *Ziel: ein Tiff pro ppn*
   - Extraktion aller ppn (Dateiname)
   - Zusammenfügen: **tiffcp**
   - **createbook.pl**
3. **FineReader**-Job (manuell, local):
   - Lese OCR-Input
   - Erzeuge Doublelayer-PDF (Upload)

Alles per **Batchjob** (Autostart)

Upload

UNIVERSITÄTSBIBLIOTHEK ILMENAU

Probleme

- Transferfehler zum GBV
- Fehlender Link
- Ungültiger Link (Tippfehler u. ä.)
- PPN-Änderung: manueller FTP-Zugriff

Upload

UNIVERSITÄTSBIBLIOTHEK ILMENAU

Probleme

- Transferfehler zum GBV
- Fehlender Link
- Ungültiger Link (Tippfehler u. ä.)
- PPN-Änderung: manueller FTP-Zugriff

FTP: ftp2GBV.pl

- Kompatibel
- Automatisch
- Einfach

Upload

UNIVERSITÄTSBIBLIOTHEK ILMENAU

Probleme

- Transferfehler zum GBV
- Fehlender Link
- Ungültiger Link (Tippfehler u. ä.)
- PPN-Änderung: manueller FTP-Zugriff

FTP: ftp2GBV.pl

- Kompatibel
- Automatisch
- Einfach

Linkcheck: checkcatalog.pl

- Tippfehler
- Transferfehler
- Falsche TOC

Upload

UNIVERSITÄTSBIBLIOTHEK ILMENAU

Probleme

- Transferfehler zum GBV
- Fehlender Link
- Ungültiger Link (Tippfehler u. ä.)
- PPN-Änderung: manueller FTP-Zugriff

FTP: ftp2GBV.pl

- Kompatibel
- Automatisch
- Einfach

Linkcheck: checkcatalog.pl

- Tippfehler
- Transferfehler
- Falsche TOC

Beide Schritte laufen permanent im Hintergrund

Upload/Check: Arbeitsschritte

UNIVERSITÄTSBIBLIOTHEK ILMENAU

1. Upload-Verzeichnis (local)

Upload/Check: Arbeitsschritte

UNIVERSITÄTSBIBLIOTHEK ILMENAU

1. Upload-Verzeichnis (local)
2. FTP-Transfer zum GBV

Upload/Check: Arbeitsschritte

UNIVERSITÄTSBIBLIOTHEK ILMENAU

1. Upload-Verzeichnis (local)
2. FTP-Transfer zum GBV
3. Verschieben nach Check (local)

Upload/Check: Arbeitsschritte

UNIVERSITÄTSBIBLIOTHEK ILMENAU

1. Upload-Verzeichnis **(local)**
2. FTP-Transfer zum GBV
3. Verschieben nach Check **(local)**
4. Abfrage des OPAC **(LWP::UserAgent)**

Upload/Check: Arbeitsschritte

UNIVERSITÄTSBIBLIOTHEK ILMENAU

1. Upload-Verzeichnis **(local)**
2. FTP-Transfer zum GBV
3. Verschieben nach Check **(local)**
4. Abfrage des OPAC **(LWP::UserAgent)**
5. TOC-Link **(regexp)**

Upload/Check: Arbeitsschritte

UNIVERSITÄTSBIBLIOTHEK ILMENAU

1. Upload-Verzeichnis (local)
2. FTP-Transfer zum GBV
3. Verschieben nach Check (local)
4. Abfrage des OPAC (LWP::UserAgent)
5. TOC-Link (regexp)
6. Download (TOC-PDF, LWP::UserAgent)

Upload/Check: Arbeitsschritte

UNIVERSITÄTSBIBLIOTHEK ILMENAU

1. Upload-Verzeichnis (local)
2. FTP-Transfer zum GBV
3. Verschieben nach Check (local)
4. Abfrage des OPAC (LWP::UserAgent)
5. TOC-Link (regexp)
6. Download (TOC-PDF, LWP::UserAgent)
7. Bitvergleich (Dowload und lokale Kopie, File::Compare)

Upload/Check: Arbeitsschritte

UNIVERSITÄTSBIBLIOTHEK ILMENAU

1. Upload-Verzeichnis (local)
2. FTP-Transfer zum GBV
3. Verschieben nach Check (local)
4. Abfrage des OPAC (**LWP::UserAgent**)
5. TOC-Link (regexp)
6. Download (TOC-PDF, **LWP::UserAgent**)
7. Bitvergleich (Dowload und lokale Kopie, **File::Compare**)
   - OK: verschiebe von Check nach Backup (LAN)
   - Fehler: verschiebe von Check nach Upload (local)
   - Falsche/Fehlende Links: Error-Log

Backup

UNIVERSITÄTSBIBLIOTHEK ILMENAU

Archivierung:

- Tiff-Input (multipage)
- Korrekt übertragene PDFs

Backup

UNIVERSITÄTSBIBLIOTHEK ILMENAU

Archivierung:

- Tiff-Input (multipage)
- Korrekt übertragene PDFs

Archivhandling:

- Archivierung: 2 Wochen
- **cleanup.pl**

Zusammenfassung

UNIVERSITÄTSBIBLIOTHEK ILMENAU

Komplettes Setup seit ca. 1 Jahr im Einsatz

Zusammenfassung

UNIVERSITÄTSBIBLIOTHEK ILMENAU

Komplettes Setup seit ca. 1 Jahr im Einsatz

- ✓ Beschleungigung des Arbeitsprozesses
- ✓ Gute Akzeptanz bei den Mitarbeiterinnen
- ✓ Gute Auslastung des OCR-PC
- ✓ Niedriger Wartungsaufwand
- ✓ Komponentennutzung auch in anderen Bereichen
- ✓ **PicaPM**: gute Perl-SRU-Schnittstelle für CBS/LBS